THE UNIVERSITY OF SYDNEY

School of Computer Science
Uwe Roehm

DATA2x01: Data Science, Big Data and Data Variety

Practical Assignment: Sydney Liveability Analysis

Group Assignment (20%) Due: 20th May 2022 @ 8 pm

Introduction

In this assignment, you will be tasked to gather and integrate several datasets in order to investigate the most 'liveable' suburb for a potential stakeholder to buy in Greater Sydney through calculation of a 'liveability' score. We then turn our attention to just the City of Sydney for a more focussed analysis.

***Disclaimer**: This assignment is mainly about data integration. Note that the age and varying quality of the provided data does not allow to reliably assess location liveability (as well as the definition of liveability being up to interpretation)*

Note: Please check Ed regularly for updates!

Preparation

In order to calculate the liveability score, several different data sources need to be integrated. As a starting point, we give you some census-based data which gives you input on quite a few statistics such as median income, number of dwellings etc.

The neighbourhood 'score' is expressed as a measure of several factors which we *assume* to affect the general 'liveability' of an area — population density, number of dwellings, entertainment interests etc. We will then correlate our score with the median income and rent of each area.

After calculating the score for each SA2 area in Greater Sydney, we turn our attention to the City of Sydney - this is when the creativity comes in! You are tasked with using the given SA2 Data as well as integrating at least 1, ideally 2 (or more) extra datasets from the City of Sydney Open Data Hub of which at one must be spatial in nature.

Your submission should consist of your Jupyter notebook that you used for integrating the data sets and for performing and visualising your analysis.

Provided Datasets

- `Neighbourhoods.csv` - `area_id, area_name, land_area, population, number_of_dwellings, number_of_businesses, median_annual_household_income, avg_monthly_rent, 0_4, 5_9, 10_14, 15_19`

- `BusinessStats.csv` - `area_id, area_name, number_of_businesses, accommodation_and_food_services, retail_trade, agriculture_forestry_and_fishing, health_care_and_social_assistance, public_administration_and_safety, transport_postal_and_warehousing`
- `SA2_2016_AUST.zip` - Statistical Area 2 (SA2) data from the Australian Bureau of Statistics (ABS) and associated parent areas.
- `break_and_enter.zip` - shape data of theft 'hotspots' in NSW as determined by BOCSAR.
- `school_catchments.zip` - contains shape data for primary, secondary and future Government schools catchments (i.e the areas in which students must live to attend each school)

Assignment Workflow

Task 1: Build a database using PostgreSQL that integrates data from the following sources:

- Neighbourhoods
- BusinessStats
- Catchments
- BreakAndEnter
- SA2 Shape Data
- Your at least 1 extra dataset for the City of Sydney Analysis

Milestone 1: Propose a stakeholder and suggest two possible datasets you will use from the City of Sydney Open Data Hub. Submit the Canvas Quiz by Week 11 Friday.

Task 2: Greater Sydney liveability Analysis

- Compute the 'liveability' score for all given neighbourhoods according to the following formula (with necessary adjustments for your extra dataset).

$$\text{Score} = \mathcal{S}(z_{\text{school}} + z_{\text{accomm}} + z_{\text{retail}} - z_{\text{crime}} + z_{\text{health}})$$

Where $\mathcal{S}$ is the sigmoid function, z is the normal z score (note you can choose to use an form of standardisation that you please, but please justify!) and we define 'young people' to be anyone aged 0 - 19.

Measure	Definition	Risk	Data Source
school	number of schools catchment areas per 1000 'young people'	+	`school_catchments.zip`
accom	number of accommodation and food services per 1000 people	+	`BusinessStats.csv`
retail	number of retail services per 1000 people	+	`BusinessStats.csv`
crime	sum of hotspot areas divided by total area	−	`break_and_enter.zip`
health	number of health services per 1000 people	+	`BusinessStats.csv`

- Visualise your score in an engaging way as well as summarising your results in a table.
- Create **at least one index** which is helpful for data integration or the liveability score computation and **justify why**.
- Determine whether there is a **correlation** between your score and the **median rent** of each neighbourhood and **median income** of each neighbourhood.

Task 3: City of Sydney Analysis

- Find at least 1 (ideally 2 or more) datasets from City of Sydney Open Data Hub of at which 1 must be spatial in nature. They should be loaded into the database in Task 1. Ideally these should be in another data format than covered by the given datasets, such as JSON or XML.
- Propose a stakeholder who wants to live in the City of Sydney. Write a short introduction (creative and fun, but realistic is great!) of them.
- Refine your 'liveability score' to now include your extra City of Sydney datasets as well as varying the coefficients to be tailored to your stakeholder (and justify why). Only perform analysis on SA2's which fall under the City of Sydney (note, this should be SA2's and suburbs where the SA3 is '*Sydney Inner City*'.)
- Visualise your results in an engaging way and make a recommendation!

Task 4: DATA2901 Task for Advanced Class ONLY

In addition to computing the given score with an explicit formula, use a supervised or unsupervised machine learning technique to "compute" a liveability score. Make sure you assess your model and check it's validity. Which score is better and why?

Deliverables

1. PDF Report: this should be no more than 6 pages plus an optional appendix in which you document your data integration steps and the main outcomes of your liveability analysis. Your document should contain the following:
   (a) *Dataset Description*: What are your data sources and how did you obtain and pre-process the data?
   (b) *Database Description*: Into which database schema did you integrate your data (preferable shown with a diagram)? Which index(es) did you create, and why?
   (c) *Greater Sydney Score Analysis*: Show which formula you applied to compute the liveability score per neighbourhood, and give an overview of the results through
   (d) *Correlation Analysis*: How well does your score correlate with the median rent and median income in each neighbourhood?
   (e) *City of Sydney Analysis*: Propose a stakeholder and give a brief introduction. Show how you tailored your score for their needs. Demonstrate the results on a map.
2. Jupyter Notebook which shows your entire data workflow
3. Access to database
4. Short Demo in Week 12 and 13 Tutorials.

All deliverables are due in Week 12, no later than **8pm, Friday 20th May 2022**.

Late submission penalty: -5% of the available marks per day late; minimum 0% after 5 days.

The marking rubric is on Canvas.

Please submit the source code and a soft copy of your documentation as a zip or tar file electronically in Canvas, one per each group. Name your zip archive after your group number X with the following name pattern: **data2001_assignment2022s1-group*X*.zip**

Students must retain electronic copies of their submitted assignment files and databases, as the unit coordinator may request to inspect these files before marking of an assignment is completed. If these assignment files are not made available to the unit coordinator when requested, the marking of this assignment may not proceed.

All the best!

Group member participation

This is a group assignment. The mark awarded for your assignment is conditional on you being able to explain any of your answers to your tutor or the lecturers if asked.

If members of your group do not contribute sufficiently you should alert your tutor as soon as possible. The tutor has the discretion to scale the group's mark for each member as follows, based on the outcome of the group's demo in Week 13.

Level of contribution	Proportion of final grade received
No participation or no demo.	0%
Passive member, but full understanding of the submitted work.	50%
Minor contributor to the group's submission.	75%
Major contributor to the group's submission.	100%